An Introduction to Data Analysis, Design of Experiment, and Machine Learning

Lecture 1. Where do data come from?

Muhammad A. Alam
alam@purdue.edu

Muhammad A Alam, Purdue University
Outline

• A short history of data
• An example of small data
• Small vs. Big data
• What to expect from the class
• Conclusions
A short history of data
Sensors and data

Infrared
Red
Violet
Ultra-violet

Cold
16Hz-
Heat
28kHz
Contact
Pain

Physical

Salt
Sour
Sweet
Bitter
Umami

Bio-chemical
Sensor

Camphor
Musk
Flowers
Mint
Ether
Acrid
Putrid

Electron microscope

Radio telescope

50e6@
2.5cm2

10k@30cm2
~3 mm2.
Outline

- A short history of data
- An example of small data
- Small vs. Big data
- What to expect from the class
- Conclusions
Time dependent dielectric breakdown

Gate current vs. log (time)

Pey, IRPS, 2002

Muhammad A Alam, Purdue University
Voltage-dependence of Dielectric Breakdown

Nonlinear voltage dependence

Empirical projection.

$V_3 > V_2 > V_1$

$V_{op}$

$\log \text{(Time to breakdown)}$

$V_G$

Muhammad A Alam, Purdue University
Weibull Distributed Failure times

Average lifetime is not good enough ....

Weibull distribution

Muhammad A Alam, Purdue University
Predictions based on data

\[ V_{P,Ox}, V_{safe} \] (volts)

\[ 0.0 \quad 0.5 \quad 1.0 \quad 1.5 \quad 2.0 \quad 2.5 \quad 3.0 \]

\[ 0.0 \quad 1.0 \quad 2.0 \quad 3.0 \quad 4.0 \quad 5.0 \]

Oxide Thickness (nm)


NMOS, PMOS

Muhammad A Alam, Purdue University
Issues with small data

- Small errors can have serious consequences.

- Generation of data is costly in terms of equipment, time, deadlines. Have to maximize information from small dataset.

- Often the dataset may be incomplete, the quality of the data non-uniform, and still we have to make the best decision possible.

- Often there could be competing hypothesis for a given distribution. Have to decide which one fits the data best. Based on the principles of Statistical decision theory.
Outline

• A short history of data
• An example of small data
• Small vs. Big data
• What to expect from the class
• Conclusions
Big vs. small data

• Big data is obtained as is. One must ask intelligent questions to tease-out the answers embedded within the information. Census and insurance information are examples. Analysis is difficult, but they do represent real world conditions.

• Small data is often hypothesis driven and obtained from carefully designed experiments or survey. Data acquisition is planned and therefore expensive. The analysis is simpler, but may not represent real world conditions.

Muhammad A Alam, Purdue University
Small vs. big data

Galileo first sketch
1610

Better telescope
1616

Published etch
1623

Muhammad A Alam, Purdue University
Where do data come from?

- Hundreds of petabyte of data every day.
- Social media sites
- Digital pictures
- Videos
- Purchase transaction
- GPS signals and so on.
- Scientific instrumentation
- Census data
Repository of big data

- Google trends
- Federal Reserve Economic Data (FRED)
- Data.gov
- US Census Bureau
- European Union Open Data Portal
- Data.gov.uk
- The CIA World Factbook
- Healthdata.gov
- NHS Health and Social Care Information Centre
- Amazon Web Services public datasets
- Facebook Graph
- Gapminder
- Google Trends
- Google Finance
- Google Books Ngrams
- National Climatic Data Center
- DBPedia
- Topsy
- Likebutton
- New York Times
- Freebase
- Million Song Data Set
- DataScienceCentral selection of big data sets - check out the first itemized bullet list after clicking on this link
- Data sets used in our data science apprenticeship - includes both real data and simulated data - and tips to create artificial, rich, big data sets for testing models
- KDNuggets repository
- Data sets used in Kaggle competitions

Muhammad A Alam, Purdue University
driven by memory technology

- Cisco estimates: 1.8 ZB by 2016 and 7.2 ZB in 2021.
- If 1 MB is the size of the period at the end of sentence, 1.8 ZB is 460 km^2, eight times the size of Manhattan
- Amazon Web services, Google Cloud, IBM Cloud, Microsoft Azure.

### Solid State Drive
- Access time: 50/1000 ns
- Capacity: 2 terabytes
- Data persistence: 8-10 years
- Read/Write Cycles: 1000

### Hard-Disk Drive
- Access time: 7 millisecond
- Capacity: 8 terabytes
- Data persistence: 3-6 years
- Read/write cycles: Indefinite

### Magnetic Tape
- Capacity: 12 terabytes
- Data persistence: 10-30 years
- Read/write cycles: Indefinite

---

Muhammad A Alam, Purdue University
Outline

• A short history of data
• An example of small data
• Small vs. Big data
• What to expect from the class
• Conclusions
What to expect ....

• A deep understanding about how to analyze the data carefully, how to fit them to analytical functions, and how to use the data to make projections.

• Overfitting of the data is a general concern. Better fitting does not imply better decisions. You will be able to recognize and exclude overfitting.

• You will learn to design the experiments and simulations systematically. And then analyze the data and understand the correlation among various inputs systematically.

• The course will introduce you to basic concepts of machine learning from a simple, intuitive perspective. It will allow you to use more powerful tools currently available.

• This is however not a course on data-science or machine learning. If you are interested, you will take online courses.

• We will take a short quiz at the end of each class to make sure that the concepts are clear.
Outline

• Course Introduction
• Collecting and Plotting Data: Robust Data Analysis
• Physical and Empirical Distribution
• Model Selection and Goodness of Fit
• Design of Experiments: Scaling of Equations
• Design of Experiments: Buckingham Pi Theorem
• Statistical Theory of Design of Experiments
• Analysis of Data: ANOVA
• Big Data Classification: Singular Value Decomposition
• Machine Learning: Part 1
• Machine Learning: Part 2
• Physics-based Machine Learning
• Course Summary, Homework and Solutions
Outline of the course

\[ \bar{y} = f(\bar{x}) \quad \bar{x} = x_1, x_2, \ldots x_n \quad \bar{y} = y_1, y_2, \ldots y_m \]

Introduction

Collecting and plotting \( x_1, x_2, \ldots x_n \)
Physical and empirical \( f, F, df/dx, \ldots \)
Model selection between \( f_1, f_2, \ldots \)

Scaling theory with known \( f \), \( f(\bar{x}) = f(\bar{X}) \)
Scaling theory with unknown \( f \), \( \bar{x} \rightarrow X \)

Principle component analysis for classifying \{y\}.
Design of experiments to determine \( \bar{y}_{\text{max}} = f(\bar{x}) \)
Machine learning … Statistical approach to learn \( f \)
Physics-based machine learning \( f = f_{\text{physics}} + \Delta f \)

Conclusions
Reference Books


Few other information

Who should take this course
Anyone interested in modeling, simulation, collecting and analyzing the data, even reading a newspaper, etc.

What are the pre-requisites
Freshman/sophomore level preparation in physics and mathematics.

Grading
Class quizzes, homeworks, one final exam.
Conclusions

• To convert data into information, we must carefully process the data, with a nuanced understanding of the implications of data processing.

• Statistical data processing techniques have dramatically over the years. A deep understanding of discrete data analysis, information-theory based curve fitting, design of experiments, machine learning, etc. will maximize the information to data ratio.